Can AI Accurately Analyze UX Survey Data?

Evaluating the Accuracy of AI Analyses Against Verified Survey Results

Research Plan









By Theresa Wilkinson

Study Objectives

Evaluate how accurately generative AI tools analyze the same UX survey dataset. Compare their calculations, grouping of written responses, and conclusions with the researcher's verified analysis to identify errors, omissions, and the human verification required.

The source dataset contains responses from 34 UX professionals. Collecting and reporting those survey responses is the focus of Study #4. This study evaluates the AI tools' analyses of those responses.

Research Questions

- Can AI accurately count responses and calculate percentages for survey questions?
- Can AI correctly count selections when respondents can choose more than one answer?
- Can AI correctly count how many respondents selected each trust rating for each of the nine activities in Question 5?
- Can AI identify, group, and count distinct ideas in written responses?
- Are the AI tools' conclusions supported by the source data?
- What errors and omissions require human verification?
- How does performance differ across the tools?

Tools and Test Setup

The original comparison evaluates:

- ChatGPT
- Claude
- Gemini
- Microsoft Copilot

A follow-up comparison evaluates:

- Dovetail
- Perplexity

Each tool receives the same survey dataset and analysis prompt. The researcher's verified findings are withheld until the tool completes its analysis.

The original outputs are retained for comparison with the source data. The four-tool comparison and the Dovetail/Perplexity comparison are reported in separate results decks.

Source Data

The dataset contains 34 responses to a 10-question survey covering:

- Professional role and UX/research experience.
- Frequency of generative AI use.
- AI tools used for UX research work.
- Trust in AI across nine UX research activities.
- Perceived need for human review.
- Comfort presenting AI-generated findings after review.
- Experience finding AI-generated errors or unsupported conclusions.
- Written explanations of which research activities respondents trust AI with most and least.

The dataset's exact question wording, response labels, and recorded answers are preserved.

Study Prompt

Analyze the attached spreadsheet exactly as provided. Do not change, clean, standardize, recode, or reinterpret any responses.

For each survey question, calculate the count and percentage for every response. For questions allowing multiple selections, calculate how many respondents selected each option and the percentage of respondents selecting it. For rating questions, report the count and percentage for each rating and calculate appropriate descriptive statistics.

Show all calculations clearly enough to verify. If any value or label is unclear or inconsistent, identify it and ask for clarification. Do not guess, correct it, or exclude it. Do not make assumptions beyond the data.

Procedure

1. Establish verified survey results through direct analysis of the source data.
2. Provide each AI tool with the same dataset and prompt.
3. Save each tool's original response before requesting corrections or clarification.
4. Compare reported counts, percentages, and descriptive statistics with the verified results.
5. Check each of Question 5's nine activities separately.
6. Review how each tool groups and counts written responses.
7. Check whether conclusions follow from the reported data.
8. Record errors, omissions, unclear reporting, and any follow-up needed to verify the output.

For written responses, distinguish between the number of respondents and the number of distinct ideas or mentions. A response may contain more than one idea.

Evaluation Criteria

- **Counting accuracy -** Do reported counts match the source data?
- **Percentage accuracy -** Are percentages calculated correctly using the appropriate denominator?
- **Multiple-selection questions -** Does the tool count each selected option correctly without treating the options as mutually exclusive?

- **Trust-rating accuracy -** Does the tool accurately report the distribution for each of Question 5's nine activities?
- **Written-response analysis -** Does the tool produce understandable categories supported by the responses? Does it distinguish respondents from distinct ideas and count them consistently?
- **Completeness -** Does the tool address each requested question and identify any analysis it could not complete?
- **Evidence and interpretation -** Are conclusions supported by the data? Are descriptive observations distinguished from statistically supported conclusions?
- **Traceability -** Is the output organized clearly enough to check each finding against its source?
- **Human verification required -** What calculations, categories, or conclusions need checking or correction before the findings can be used?

Analysis and Reporting

Evaluate each tool against the verified source results using the same criteria. Report accurate findings, errors, omissions, and incomplete analyses separately.

Identify the specific questions or activities affected by each error. Correct results on some questions will not be treated as evidence that the entire analysis is accurate.

For written responses, assess whether categories reflect the actual responses and whether the counting approach matches the intended unit of analysis. Different category wording alone will not establish an error.

Report the original four-tool comparison separately from the Dovetail/Perplexity follow-up.

Additional statistical analyses generated by Dovetail and Perplexity are outside the evaluated scope. Any question or output not reviewed in detail will be identified as such.

Limitations

- The study evaluates performance on one survey dataset containing 34 responses.
- Results apply to the specific prompts and outputs evaluated; they do not establish how a tool will perform on every dataset.
- AI outputs may vary across repeated runs and model updates.

- The evaluation is conducted by one experienced UX researcher.
- Written-response categorization involves judgment. Alternative categories may be reasonable when supported by the responses and applied consistently.
- The study does not establish that one tool is universally more accurate than another.
- Human verification effort is described qualitatively unless time was measured systematically.
- The study evaluates analysis of existing survey data. It does not evaluate recruiting, moderation, or data collection.